On Hidden Markov Processes with Infinite Excess Entropy

Łukasz Dębowski*

Abstract

We investigate stationary hidden Markov processes for which mutual information between the past and the future is infinite. It is assumed that the number of observable states is finite and the number of hidden states is countably infinite. Under this assumption, we show that the block mutual information of a hidden Markov process is upper bounded by a power law determined by the tail index of the hidden state distribution. Moreover, we exhibit three examples of processes. The first example, considered previously, is nonergodic and the mutual information between the blocks is bounded by the logarithm of the block length. The second example is also nonergodic but the mutual information between the blocks obeys a power law. The third example obeys the power law and is ergodic.

Key words: hidden Markov processes, mutual information, ergodic processes

MSC 2010: 60J10, 94A17, 37A25

Running head: On Processes with Infinite Excess Entropy

*Ł. Dębowski is with the Institute of Computer Science, Polish Academy of Sciences, ul. Jana Kazimierza 5, 01-248 Warszawa, Poland (e-mail: ldebowsk@ipipan.waw.pl).

1 Introduction

In recent years there has been a surge of interdisciplinary interest in excess entropy, which is the Shannon mutual information between the past and the future of a stationary discrete-time process. The initial motivation for this interest was a paper by Hilberg [22], who supposed that certain processes with infinite excess entropy may be useful for modeling texts in natural language. Subsequently, it was noticed that processes with infinite excess entropy appear also in research of other, so called, complex systems [6, 13, 14, 19, 1, 25, 5, 23]. Also from a purely mathematical point of view, excess entropy is an interesting measure of dependence for nominal valued random processes, where the analysis of autocorrelation does not provide sufficient insight into process memory.

Briefly reviewing earlier works, let us mention that excess entropy has been already studied for several classes of processes. The most classical results concern Gaussian processes, where Grenander and Szegő [20, Section 5.5] gave an integral formula for excess entropy (in disguise) and Finch [18] evaluated this formula for autoregressive moving average (ARMA) processes. In the ARMA case excess entropy is finite. A few more papers concern processes over a finite alphabet with infinite excess entropy. For instance, Bradley [3] constructed the first example of a mixing process having this property. Gramss [19] investigated a process which is formed by the frequencies of 0's and 1's in the rabbit sequence. Travers and Crutchfield [26] researched some hidden Markov processes with a countably infinite number of hidden states. Some attempts were also made to generalize excess entropy to two-dimensional random fields [17, 4].

Excess entropy is an intuitive measure of memory stored in a stochastic process. Although this quantity only measures the memory capacity, without characterizing *how* the process future depends on the past, it can be given interesting general interpretations. Mahoney, Ellison and Crutchfield [24, 15] developed a formula for excess entropy in terms of predictive and retrodictive ϵ-machines, which are minimal unifilar hidden Markov representations of the process [25, 23]. In our previous works [10, 11, 12, 9], we also investigated excess entropy of stationary processes that model texts in natural language. We showed that a power-law growth of mutual information between adjacent blocks of text arises when the text describes certain facts in a logically consistent and highly repetitive way. Moreover, if the mutual information between blocks grows according to a power law then a similar power law is obeyed by the number of distinct words, identified formally as codewords in a certain text compression [7]. The latter power law is known as Herdan's law [21], which is an integral version of the famous Zipf law observed for natural language [28].

In this paper we will study several examples of stationary hidden Markov processes over a finite alphabet for which excess entropy is infinite. The first study of such processes was developed by Travers and Crutchfield [26]. A few more words about the adopted setting are in need. First, excess entropy is finite for hidden Markov chains with a finite number of hidden states. This is the usually studied case [16], for which the name of finite-state sources is also used. To allow for hidden Markov processes with unbounded mutual information, we need to assume that the number of hidden states is at least countably infinite. Second, we want to restrict the class of studied models. If we admitted an uncountable number of hidden states or a nonstationary distribution over the hidden states then the class of hidden Markov processes would cover all

processes (over a countable alphabet). For that reason we will assume that the underlying Markov process is stationary and the number of hidden states is exactly countably infinite. In contrast, the number of observable states is fixed as finite to focus on nontrivial examples. In all these assumptions we follow [26].

The modest aim of the present paper is to demonstrate that power-law growth of mutual information between adjacent blocks may arise for very simple hidden Markov processes. Presumably, stochastic processes which exhibit this power law appear in modeling of natural language [22, 10]. But the processes that we study here do not have a clear linguistic interpretation. They are only mathematical instances presented to show what is possible in theory. Although these processes are simple to define, we perceive them as somehow artificial because of the way *how* the memory of the past is stored in the present and revealed in the future. Understanding what are acceptable mechanisms of memory in realistic stochastic models of complex systems is an important challenge for future research.

The further organization of the paper is as follows: In Section 2 we present the results, whereas the proofs are deferred to Section 3.

2 Results

Now we begin the formal presentation of our results. First, let $(Y_i)_{i\in\mathbb{Z}}$ be a stationary Markov process on $(\Omega, \mathcal{J}, P)$ where variables $Y_i : \Omega \to \mathbb{Y}$ take values in a countably infinite alphabet $\mathbb{Y}$. This process is called the hidden process. Next, for a function $f : \mathbb{Y} \to \mathbb{X}$, where the alphabet $\mathbb{X} = \{0, 1, ..., D-1\}$ is finite, we construct process $(X_i)_{i\in\mathbb{Z}}$ with

$$X_i = f(Y_i). \tag{1}$$

Process $(X_i)_{i\in\mathbb{Z}}$ will be called the observable process. The process is called unifilar if $Y_{i+1} = g(Y_i, X_{i+1})$ for a certain function $g : \mathbb{Y} \times \mathbb{X} \to \mathbb{Y}$. Such a construction of hidden Markov processes, historically the oldest one [2], is called state-emitting (or Moore) in contrast to another construction named edge-emitting (or Mealy). The Mealy construction, with a requirement of unifilarity, has been adopted in previous works [5, 26, 23]. Here, we adopt the Moore construction and we drop the unifilarity assumption since it leads to a simpler presentation of processes. It should be noted that the standard definition of hidden Markov processes in statistics and signal processing is yet up to a degree different, namely the observed process $(X_i)_{i\in\mathbb{Z}}$ depends on the hidden process $(Y_i)_{i\in\mathbb{Z}}$ via a probability distribution and X_i is conditionally independent of the other observables given Y_i. All the presented definitions are, however, equivalent and the terminological discussion can be put aside.

In the following turn we inspect the mutual information. Having entropy $H(X) = \mathbf{E}\left[-\log P(X)\right]$ with log denoting the binary logarithm throughout this paper, mutual information is defined as $I(X;Y) = H(X) + H(Y) - H(X,Y)$. Here we will be interested in the block mutual information of the observable process,

$$E(n) := I(X^0_{-n+1}; X^n_1), \tag{2}$$

where X^l_k denotes the block $(X_i)_{k\le i\le l}$. More specifically, we are interested in processes for which excess entropy $E = \lim_{n\to\infty} E(n)$ is infinite and $E(n)$

diverges at a power-law rate. We want to show that such an effect is possible for very simple hidden Markov processes. (Travers and Crutchfield [26] considered some examples of nonergodic and ergodic hidden Markov processes with infinite excess entropy but they did not investigate the rate of divergence of $E(n)$.) Notice that by the data processing inequality for the Markov process $(Y_i)_{i\in\mathbb{Z}}$, we have

$$E(n) \leq I(Y_{-n+1}^0; Y_1^n) = I(Y_0; Y_1) \leq H(Y_0). \tag{3}$$

Thus the block mutual information $E(n)$ may diverge only if the entropy of the hidden state is infinite. To achieve this effect, the hidden variable Y_0 must necessarily assume an infinite number of values.

Now we introduce our class of examples. Let us assume that hidden states σ_{nk} may be grouped into levels

$$T_n := \{\sigma_{nk}\}_{1\leq k\leq r(n)} \tag{4}$$

that comprise equiprobable values. Moreover, we suppose that the level indicator

$$N_i := n \iff Y_i \in T_n \tag{5}$$

has distribution

$$P(N_i = n) = \frac{C}{n \log^\alpha n}. \tag{6}$$

For $\alpha \in (1, 2]$, entropy $H(N_i)$ is infinite and so is $H(Y_i) \geq H(N_i)$ since N_i is a function of Y_i. In the following, we work with this specific distribution of Y_i.

As we will show, the rate of growth of the block mutual information $E(n)$ is bounded in terms of exponent α from equation (6). Let us write $f(n) = O(g(n))$ if $f(n) \leq Kg(n)$ for a $K > 0$ and $f(n) = \Theta(g(n))$ if $K_1 g(n) \leq f(n) \leq K_2 g(n)$ for $K_1, K_2 > 0$.

Theorem 1 *Assume that* $\mathbb{Y} = \{\sigma_{nk}\}_{1\leq k\leq r(n), n\geq 2}$, *where function* $r(n)$ *satisfies* $r(n) = O(n^p)$ *for a* $p \in \mathbb{N}$. *Moreover assume that*

$$P(Y_i = \sigma_{nk}) = \frac{1}{r(n)} \cdot \frac{C}{n \log^\alpha n}, \tag{7}$$

where $\alpha \in (1, 2]$ *and* $C^{-1} = \sum_{n=2}^{\infty} (n \log^\alpha n)^{-1}$. *Then we have*

$$E(n) = \begin{cases} O\left(n^{2-\alpha}\right), & \alpha \in (1, 2), \\ O\left(\log n\right), & \alpha = 2. \end{cases} \tag{8}$$

The interesting question becomes whether there exist hidden Markov processes that achieve the upper bound established in Theorem 1. If so, can they be ergodic? The answer to both questions is positive and we will exhibit some simple examples of such processes.

The first example that we present is nonergodic and the mutual information diverges slower than expected from Theorem 1.

Example 1 (Heavy Tailed Periodic Mixture I) *This example has been introduced in [26]. We assume* $\mathbb{Y} = \{\sigma_{nk}\}_{1\le k\le r(n), n\ge 2}$, *where* $r(n) = n$. *Then we set the transition probabilities*

$$P(Y_{i+1} = \sigma_{nk} | Y_i = \sigma_{ml}) = \begin{cases} \mathbf{1}\{n = m, k = l+1\}, & 1 \le l \le m-1, \\ \mathbf{1}\{n = m, k = 1\}, & l = m. \end{cases} \tag{9}$$

We can see that the transition graph associated with the process $(Y_i)_{i\in\mathbb{Z}}$ *consists of disjoint cycles on levels* T_n. *The stationary distribution of the Markov process is not unique and the process is nonergodic if more than one cycle has a positive probability. Here we assume the cycle distribution (6) so the stationary marginal distribution of* Y_i *equals (7). Moreover, the observable process is set as*

$$X_i = \begin{cases} 0, & Y_i = \sigma_{nk},\ 1 \le k \le n-1, \\ 1, & Y_i = \sigma_{nn}. \end{cases} \tag{10}$$

In the above example, the level indicator N_i has infinite entropy and is measurable with respect to the shift invariant algebra of the observable process $(X_i)_{i\in\mathbb{Z}}$. Hence $E(n)$ tends to infinity by the ergodic decomposition of excess entropy [8, Theorem 5]. A more precise bound on the block mutual information is given below.

Proposition 1 *For Example 1, we have*

$$E(n) = \begin{cases} \Theta(\log^{2-\alpha} n), & \alpha \in (1,2), \\ \Theta(\log\log n), & \alpha = 2. \end{cases} \tag{11}$$

The next example is also nonergodic but the rate of mutual information reaches the upper bound. It seems to happen so because the information about the hidden state level is coded in the observable process in a more concise way.

Example 2 (Heavy Tailed Periodic Mixture II) *We assume that* $\mathbb{Y} = \{\sigma_{nk}\}_{1\le k\le r(n), n\ge 2}$, *where* $r(n) = s(n)$ *is the length of the binary expansion of number* n. *Then we set the transition probabilities*

$$P(Y_{i+1} = \sigma_{nk} | Y_i = \sigma_{ml}) = \begin{cases} \mathbf{1}\{n = m, k = l+1\}, & 1 \le l \le s(m)-1, \\ \mathbf{1}\{n = m, k = 1\}, & l = s(m). \end{cases} \tag{12}$$

Again, the transition graph associated with the process $(Y_i)_{i\in\mathbb{Z}}$ *consists of disjoint cycles on levels* T_n. *As previously, we assume the cycle distribution (6) and the marginal distribution (7). Moreover, let* $b(n,k)$ *be the* k*-th digit of the binary expansion of number* n. *(We have* $b(n,1) = 1$.*) The observable process is set as*

$$X_i = \begin{cases} 2, & Y_i = \sigma_{n1}, \\ b(n,k), & Y_i = \sigma_{nk}, 2 \le k \le s(n). \end{cases} \tag{13}$$

Proposition 2 *For Example 2, we have*

$$E(n) = \begin{cases} \Theta(n^{2-\alpha}), & \alpha \in (1,2), \\ \Theta(\log n), & \alpha = 2. \end{cases} \tag{14}$$

In the third example the rate of mutual information also reaches the upper bound and the process is additionally ergodic. The process resembles the Branching Copy (BC) process introduced in [26]. There are three main differences between the BC process and our process. First, we discuss a simpler nonunifilar presentation of the process rather than a more complicated unifilar one. Second, we add strings of separators $(s(m)+1)\times 3$ in the observable process. Third, we put slightly different transition probabilities to obtain a simpler stationary distribution. All these changes lead to a simpler computation of mutual information.

Example 3 (Heavy Tailed Mixing Copy) *Let $\mathbb{Y}=\{\sigma_{nk}\}_{1\le k\le r(n),n\ge 2}$ with $r(n)=3s(n)$ and $s(n)$ being the length of the binary expansion of number n. Then we set the transition probabilities*

$$P(Y_{i+1}=\sigma_{nk}|Y_i=\sigma_{ml})=\begin{cases}\mathbf{1}\{n=m,k=l+1\}, & 1\le l\le r(m)-1,\\ p(n)\mathbf{1}\{k=1\}, & l=r(m),\end{cases} \tag{15}$$

where

$$p(n)=\frac{1}{r(n)}\cdot\frac{D}{n\log^{\alpha}n} \tag{16}$$

and $D^{-1}=\sum_{n=2}^{\infty}(r(n)\cdot n\log^{\alpha}n)^{-1}$. This time levels T_n communicate through transitions $\sigma_{mr(m)}\to\sigma_{n1}$ happening with probabilities $p(n)$. The transition graph of the process $(Y_i)_{i\in\mathbb{Z}}$ is strongly connected and there is a unique stationary distribution. Hence the process is ergodic. It can be easily verified that the stationary distribution is (7) so the levels are distributed according to (6). As previously, let $b(n,k)$ be the k-th digit of the binary expansion of number n. The observable process is set as

$$X_i=\begin{cases}2, & Y_i=\sigma_{n1},\\ b(n,k), & Y_i=\sigma_{nk},\ 2\le k\le s(n),\\ 3, & Y_i=\sigma_{nk},\ s(n)+1\le k\le 2s(n)+1,\\ b(n,k-2s(n)), & Y_i=\sigma_{nk},\ 2s(n)+2\le k\le 3s(n).\end{cases} \tag{17}$$

Proposition 3 *For Example 3, $E(n)$ satisfies (14).*

Resuming our results, we make this comment. The power-law growth of block mutual information has been previously considered a hallmark of stochastic processes that model "complex behavior", such as texts in natural language [22, 1, 5]. However, the constructed examples of hidden Markov processes feature quite simple transition graphs. Consequently, one may doubt whether power-law growth of mutual information is a sufficient reason to call a given stochastic process a model of complex behavior, even when we restrict the class of processes to processes over a finite alphabet. Basing on our experience with other processes with rapidly growing block mutual information [10, 11, 12, 9], which are more motivated linguistically, we think that infinite excess entropy is just one of the necessary conditions. Identifying other conditions for stochastic models of complex systems is a matter of further interdisciplinary research. We believe that these conditions depend on a particular system to be modeled.

3 Proofs

We begin with two simple bounds.

Lemma 1 *Let $\alpha \in (1,2]$. On the one hand, we have*

$$\sum_{m=2}^{n} \frac{1}{m \log^{\alpha-1} m} = \delta_1 + \begin{cases} \frac{\ln 2}{2-\alpha} \left(\log^{2-\alpha} n - 1 \right), & \alpha \in (1,2), \\ (\ln^2 2) \log\log n, & \alpha = 2, \end{cases} \tag{18}$$

where $0 \le \delta_1 \le 1/2$. On the other hand, we have

$$\sum_{m=n}^{\infty} \frac{1}{m \log^{\alpha} m} = \delta_2 + \frac{\ln 2}{\alpha - 1} \log^{1-\alpha} n \tag{19}$$

where $0 \le \delta_2 \le (n \log^{\alpha} n)^{-1}$.

Proof: For a continuous decreasing function f we have $\int_a^b f(m)dm \le \sum_{m=a}^b f(m) \le f(a) + \int_a^b f(m)dm$. Moreover,

$$\int_2^n \frac{dm}{m \log^{\alpha-1} m} = \int_1^{\log n} \frac{(\ln 2)dp}{p^{\alpha-1}} = \begin{cases} \frac{\ln 2}{2-\alpha} \left(\log^{2-\alpha} n - 1 \right), & \alpha \in (1,2), \\ (\ln^2 2) \log\log n, & \alpha = 2, \end{cases}$$

$$\int_n^{\infty} \frac{dm}{m \log^{\alpha} m} = \int_{\log n}^{\infty} \frac{(\ln 2)dp}{p^{\alpha}} = \frac{\ln 2}{\alpha - 1} \log^{1-\alpha} n.$$

Hence the claims follow. □

For an event B, let us introduce conditional entropy $H(X|B)$ and mutual information $I(X;Y|B)$, which are respectively the entropy of variable X and mutual information between variables X and Y taken with respect to probability measure $P(\cdot|B)$. The conditional entropy $H(X|Z)$ and information $I(X;Y|Z)$ for a variable Z are the averages of expressions $H(X|Z=z)$ and $I(X;Y|Z=z)$ taken with weights $P(Z=z)$. That is the received knowledge. Now comes a handy fact that we will also use. Let I_B be the indicator function of event B. Observe that

$$\begin{aligned} I(X;Y) &= I(X;Y|I_B) + I(X;Y;I_B) \\ &= P(B) I(X;Y|B) + P(B^c) I(X;Y|B^c) + I(X;Y;I_B), \end{aligned} \tag{20}$$

where the triple information $I(X;Y;I_B)$ satisfies $|I(X;Y;I_B)| \le H(I_B) \le 1$ by the information diagram [27].

Proof of Theorem 1: Consider the event $B = (N_o \le 2^n)$, where N_0 is the level indicator of variable Y_0. On the one hand, by Markovianity of $(Y_i)_{i \in \mathbb{Z}}$, we have

$$\begin{aligned} I(X_{-n+1}^0; X_1^n | B) &\le I(Y_{-n+1}^0; Y_1^n | B) \\ &\le I(Y_0; Y_1 | B) \le H(Y_0 | B). \end{aligned}$$

On the other hand, for B^c, the complement of B, we have

$$I(X_{-n+1}^0; X_1^n | B^c) \le H(X_{-n+1}^0 | B^c) \le n \log |\mathbb{X}|,$$

where $|\mathbb{X}|$, the cardinality of set $\mathbb{X}$, is finite. Hence, using (20), we obtain

$$\begin{aligned} E(n) &\le P(B) I(X_{-n+1}^0; X_1^n|B) + P(B^c) I(X_{-n+1}^0; X_1^n|B^c) + 1 \\ &\le P(B) H(Y_0|B) + nP(B^c) \log |\mathbb{X}| + 1, \end{aligned} \tag{21}$$

where

$$P(B) = \sum_{m=2}^{2^n} \frac{C}{m \log^\alpha m}.$$

Using (18) yields further

$$\begin{aligned} P(B)H(Y_0|B) &= P(B) \sum_{m=2}^{2^n} \frac{C}{P(B) \cdot m \log^\alpha m} \log \frac{P(B) \cdot r(m) \cdot m \log^\alpha m}{C} \\ &= \sum_{m=2}^{2^n} \frac{C}{m \log^\alpha m} \log \frac{r(m) \cdot m \log^\alpha m}{C} + P(B) \log P(B) \\ &= \Theta\left(\sum_{m=2}^{2^n} \frac{1}{m \log^{\alpha-1} m} \right) = \begin{cases} \Theta\left(n^{2-\alpha}\right), & \alpha \in (1,2), \\ \Theta\left(\log n\right), & \alpha = 2. \end{cases} \end{aligned}$$

On the other hand, by (19), we have

$$nP(B^c) = n \sum_{m=2^n+1}^{\infty} \frac{C}{m \log^\alpha m} = \Theta\left(n^{2-\alpha}\right).$$

Plugging both bounds into (21) yields the requested bound (8). □

Now we prove Propositions 1–3. The proofs are very similar and consist in constructing variables D_n that are both functions of X_{-n+1}^0 and functions of X_1^n. Given this property, we obtain

$$\begin{aligned} E(n) &= I(X_{-n+1}^0, D_n; X_1^n) = I(D_n; X_1^n) + I(X_{-n+1}^0; X_1^n|D_n) \\ &= H(D_n) + I(X_{-n+1}^0; X_1^n|D_n). \end{aligned} \tag{22}$$

Hence, some lower bounds for the block mutual information $E(n)$ follow from the respective bounds for the entropies of D_n.

Proof of Proposition 1: Introduce random variable

$$D_n = \begin{cases} N_0, & 2N_0 \le n, \\ 0, & 2N_0 > n. \end{cases}$$

Equivalently, we have

$$D_n = \begin{cases} N_1, & 2N_1 \le n, \\ 0, & 2N_1 > n. \end{cases}$$

It can be seen that D_n is both a function of X_{-n+1}^0 and a function of X_1^n. On the one hand, observe that if $2N_0 \le n$ then we can identify N_0 given X_{-n+1}^0

because the full period is visible in X_{-n+1}^0, bounded by two delimiters 1. On the other hand, if $2N_0 > n$ then given X_{-n+1}^0 we may conclude that the period's length N_0 exceeds $n/2$, regardless whether the whole period is visible or not. Hence variable D_n is a function of X_{-n+1}^0. In a similar way, we show that D_n is a function of X_1^n. Given both facts, we derive (22).

Next, we bound the terms appearing on the right hand side of (22). For a given N_0, variable X_{-n+1}^0 assumes at most N_0 distinct values, which depend on N_0. Hence

$$H(X_{-n+1}^0|D_n = m) \leq \log m \text{ for } 2 \leq m \leq \lfloor n/2 \rfloor.$$

On the other hand, if we know that $N_0 > n$ then the number of distinct values of variable X_{-n+1}^0 equals $n+1$. Consequently, if we know that $D_n = 0$, i.e., $N_0 \geq \lfloor n/2 \rfloor + 1$, then the number of distinct values of X_{-n+1}^0 is bounded above by

$$\begin{aligned} n+1+\sum_{m=\lfloor n/2 \rfloor+1}^{n} m &= n+1+\frac{n(n+1)}{2}+\frac{\lfloor n/2 \rfloor\,(\lfloor n/2 \rfloor+1)}{2} \\ &\leq \frac{3n^2+14n+8}{8} \leq \frac{25}{8}n^2. \end{aligned}$$

In this way we obtain

$$H(X_{-n+1}^0|D_n = 0) \leq \log(25n^2/8).$$

Hence, by (18) and (19), the conditional mutual information may be bounded

$$\begin{aligned} I(X_{-n+1}^0; X_1^n|D_n) &\leq H(X_{-n+1}^0|D_n) \\ &= \sum_{m=2}^{\lfloor n/2 \rfloor} P(D_n = m) H(X_{-n+1}^0|D_n = m) \\ &\quad + P(D_n = 0) H(X_{-n+1}^0|D_n = 0) \\ &\leq \sum_{m=2}^{\lfloor n/2 \rfloor} \frac{C \log m}{m \log^\alpha m} + \sum_{m=\lfloor n/2 \rfloor+1}^{\infty} \frac{C \log(25n^2/8)}{m \log^\alpha m} \\ &= \begin{cases} \Theta(\log^{2-\alpha} n), & \alpha \in (1,2), \\ \Theta(\log\log n), & \alpha = 2. \end{cases} \end{aligned}$$

The entropy of D_n may be bounded similarly,

$$\begin{aligned} H(D_n) &= \sum_{m=2}^{\lfloor n/2 \rfloor} \frac{C}{m \log^\alpha m} \log \frac{m \log^\alpha m}{C} - P(D_n = 0) \log P(D_n = 0) \\ &= \begin{cases} \Theta(\log^{2-\alpha} n), & \alpha \in (1,2), \\ \Theta(\log\log n), & \alpha = 2. \end{cases} \end{aligned}$$

Hence, because $E(n)$ satisfies (22), we obtain (11). □

Proof of Proposition 2: Introduce random variable

$$D_n = \begin{cases} N_0, & 2s(N_0) \leq n, \\ 0, & 2s(N_0) > n. \end{cases}$$

Equivalently, we have

$$D_n = \begin{cases} N_1, & 2s(N_1) \le n, \\ 0, & 2s(N_1) > n. \end{cases}$$

As in the previous proof, the newly constructed variable D_n is both a function of X_{-n+1}^0 and a function of X_1^n. If $2s(N_0) \le n$ then we can identify N_0 given X_{-n+1}^0 because the full period is visible in X_{-n+1}^0, bounded by two delimiters 2. If $2s(N_0) > n$ then given X_{-n+1}^0 we may conclude that the period's length $s(N_0)$ exceeds $n/2$, regardless whether the whole period is visible or not. Hence variable D_n is a function of X_{-n+1}^0. In a similar way, we demonstrate that D_n is a function of X_1^n. By these two facts, we infer (22).

Observe that the largest m such that $s(m) = \lfloor \log m \rfloor + 1 \le \lfloor n/2 \rfloor$ is $m = 2^{\lfloor n/2 \rfloor} - 1$. Using (18), the entropy of D_n may be bounded as

$$\begin{aligned} H(D_n) &= \sum_{m=2}^{2^{\lfloor n/2 \rfloor}-1} \frac{C}{m \log^\alpha m} \log \frac{m \log^\alpha m}{C} - P(D_n = 0) \log P(D_n = 0) \\ &= \begin{cases} \Theta(n^{2-\alpha}), & \alpha \in (1,2), \\ \Theta(\log n), & \alpha = 2, \end{cases} \end{aligned}$$

Thus (14) follows by (22) and Theorem 1. □

Proof of Proposition 3: Introduce random variable

$$D_n = \begin{cases} m, & Y_0 = \sigma_{mk},\ s(m)+1 \le k \le 2s(m),\ 2s(m) \le n, \\ 0, & \text{else.} \end{cases}$$

Equivalently, we have

$$D_n = \begin{cases} m, & Y_1 = \sigma_{mk},\ s(m)+2 \le k \le 2s(m)+1,\ 2s(m) \le n, \\ 0, & \text{else.} \end{cases}$$

Again, it can be seen that D_n is both a function of X_{-n+1}^0 and a function of X_1^n. The way of computing D_n given X_{-n+1}^0 is as follows. If

$$X_{-n+1}^0 = (..., 2, b(m,2), b(m,3), ..., b(m,s(m)), \underbrace{3, ..., 3}_{l \text{ times}})$$

for some m such that $2s(m) \le n$ and $1 \le l \le s(m)$ then we return $D_n = m$. Otherwise we return $D_n = 0$. The recipe for D_n given X_1^n is mirror-like. If

$$X_1^n = (\underbrace{3, ..., 3}_{l \text{ times}}, b(m,2), b(m,3), ..., b(m,s(m)), 2, ...)$$

for some m such that $2s(m) \le n$ and $1 \le l \le s(m)$ then we return $D_n = m$. Otherwise we return $D_n = 0$. In view of these observations we derive (22), as in the previous two proofs.

Now, for $m \ne 0$ and $s(m) \le n/2$, the distribution of D_n is

$$P(D_n = m) = \frac{s(m)}{3s(m)} \cdot \frac{C}{m \log^\alpha m} = \frac{1}{3} \cdot \frac{C}{m \log^\alpha m}.$$

Notice that the largest m such that $s(m) = \lfloor \log m \rfloor + 1 \leq \lfloor n/2 \rfloor$ is $m = 2^{\lfloor n/2 \rfloor} - 1$. Hence, by (18), the bound for the entropy of D_n is

$$\begin{aligned} H(D_n) &= \sum_{m=2}^{2^{\lfloor n/2 \rfloor}-1} \frac{C}{3m \log^{\alpha} m} \log \frac{3m \log^{\alpha} m}{C} - P(D_n = 0) \log P(D_n = 0) \\ &= \begin{cases} \Theta(n^{2-\alpha}), & \alpha \in (1,2), \\ \Theta(\log n), & \alpha = 2, \end{cases} \end{aligned}$$

Consequently, (14) follows by (22) and Theorem 1. □

Acknowledgment

I thank Nick Travers, Jan Mielniczuk, and an anonymous referee for comments and remarks.

References

[1] W. Bialek, I. Nemenman, and N. Tishby. Complexity through nonextensivity. *Physica A*, 302:89–99, 2001.

[2] D. Blackwell. The entropy of functions of finite-state Markov chains. In *Transactions of the First Prague Conference on Information Theory, Statistical Decision Functions, Random Processes*, pages 13–20. Czechoslovak Academy of Sciences, 1956.

[3] R. C. Bradley. On the strong mixing and weak Bernoulli conditions. *Z. Wahrsch. verw. Geb.*, 50:49–54, 1980.

[4] W. Bułatek and B. Kamiński. On excess entropies for stationary random fields. *Probab˙ Math. Statist.*, 29:353–367, 2009.

[5] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, 15:25–54, 2003.

[6] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, 1989.

[7] C. G. de Marcken. *Unsupervised Language Acquisition*. PhD thesis, Massachussetts Institute of Technology, 1996.

[8] Ł. Dębowski. A general definition of conditional information and its application to ergodic decomposition. *Statist. Probab. Lett.*, 79:1260–1268, 2009.

[9] Ł. Dębowski. Variable-length coding of two-sided asymptotically mean stationary measures. *J. Theor. Probab.*, 23:237–256, 2010.

[10] Ł. Dębowski. On the vocabulary of grammar-based codes and the logical consistency of texts. *IEEE Trans. Inform. Theor.*, 57:4589–4599, 2011.

[11] Ł. Dębowski. Excess entropy in natural language: present state and perspectives. *Chaos*, 21:037105, 2011.

[12] Ł. Dębowski. Mixing, ergodic, and nonergodic processes with rapidly growing information between blocks. *IEEE Trans. Inform. Theor.*, 58:3392–3401, 2012.

[13] W. Ebeling. Prediction and entropy of nonlinear dynamical systems and symbolic sequences with LRO. *Physica D*, 109:42–45, 1997.

[14] W. Ebeling and T. Pöschel. Entropy and long-range correlations in literary English. *Europhys. Lett.*, 26:241–246, 1994.

[15] C. J. Ellison, J. R. Mahoney, and J. P. Crutchfield. Prediction, retrodiction, and the amount of information stored in the present. *J. Statist. Phys.*, 136: 1005–1034, 2009.

[16] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theor.*, 48:1518–1569, 2002.

[17] D. P. Feldman and J. P. Crutchfield. Structural information in two-dimensional patterns: Entropy convergence and excess entropy. *Phys. Rev. E*, 67:051104, 2003.

[18] P. D. Finch. On the covariance determinants of autoregressive and moving average models. *Biometrika*, 47:194–211, 1960.

[19] T. Gramss. Entropy of the symbolic sequence for critical circle maps. *Phys. Rev. E*, 50:2616–2620, 1994.

[20] U. Grenander and G. Szegő. *Toeplitz Forms and Their Applications*. Berkeley: University of California Press, 1958.

[21] G. Herdan. *Quantitative Linguistics*. Butterworths, 1964.

[22] W. Hilberg. Der bekannte Grenzwert der redundanzfreien Information in Texten — eine Fehlinterpretation der Shannonschen Experimente? *Frequenz*, 44:243–248, 1990.

[23] W. Löhr. Properties of the statistical complexity functional and partially deterministic HMMs. *Entropy*, 11:385–401, 2009.

[24] J. R. Mahoney, C. J. Ellison, and J. P. Crutchfield. Information accessibility and cryptic processes. *J. Phys. A*, 42:362002, 2009.

[25] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Statist. Phys.*, 104:819–881, 2001.

[26] N. F. Travers and J. P. Crutchfield. Infinite excess entropy processes with countable-state generators. `http://arxiv.org/abs/1111.3393`, 2011.

[27] R. W. Yeung. *First Course in Information Theory*. Kluwer Academic Publishers, 2002.

[28] G. K. Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin, 1935.